PSEUDOMARKER 2.0: efficient computation of likelihoods using NOMAD

E Michael Gertz*[1,*], Tero Hiekkalinna[2,3], Sébastien Le Digabel[4], Charles Audet[4], Joseph D Terwilliger[2,5,6], Alejandro A Schäffer[1]

[1]National Center for Biotechnology Information, NIH, DHHS, Bethesda, MD, USA
[2]Unit of Public Health Genomics, National Institute for Health and Welfare, Helsinki, FINLAND
[3]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, FINLAND
[4]GERAD and Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, Montréal, CANADA
[5]Department of Psychiatry, Department of Genetics and Development, and Columbia Genome Center, Columbia University, New York, NY, USA
[6]Division of Medical Genetics, New York State Psychiatric Institute, New York, NY USA

Email: E Michael Gertz*- gertz@ncbi.nlm.nih.gov; Tero Hiekkalinna - tero.hiekkalinna@helsinki.fi; Sébastien Le Digabel - Sebastien.Le.Digabel@gerad.ca; Charles Audet - Charles.Audet@gerad.ca; Joseph D Terwilliger - jdt3@columbia.edu; Alejandro A Schäffer - aschaffe@ncbi.nlm.nih.gov;

*Corresponding author

Abstract

Background: PSEUDOMARKER is a software package that performs joint linkage and linkage disequilibrium analysis between a marker and a putative disease locus.

Results: The new version 2.0 uses the software package NOMAD to maximize likelihoods, resulting in generally comparable or better optima with many fewer evaluations of the likelihood functions.

Conclusions: After being modified substantially to use modern optimization methods, PSEUDOMARKER version 2.0 is more robust and substantially faster than version 1.0. NOMAD may be useful in other bioinformatics problems where complex likelihood functions are optimized.

Background

PSEUDOMARKER [1] is a package that genomically localizes trait-predisposing loci by performing statistical tests using a putative disease locus and a series of markers. Typically, such packages either test for cosegregation by descent of markers with a putative trait-predisposing locus within families (genetic linkage), or test for differences in marker genotype frequencies between unrelated cases and controls (linkage disequilibrium (LD)). With PSEUDOMARKER one can jointly analyze linkage and LD, or LD conditional on linkage, using the pedigree relationships among individuals where known.

PSEUDOMARKER version 1 maximizes several likelihood functions [1] using a generalized pattern search (GPS) algorithm [2] implemented in a custom version of the ILINK [3] program. Previously, we showed that PSEUDOMARKER, using GPS likelihood estimates, performed well in detecting linkage and LD, outperforming several competing genetic analysis programs as measured by the power or false positive rate [4].

The running time of PSEUDOMARKER depends on the number of times the optimization algorithm evaluates any likelihood function. Each evaluation involves computation over one or, often, several pedigrees for fixed values of certain parameters that may include the recombination fraction and marker allele frequencies. ILINK computes these likelihoods using a peeling method that is a generalization of the Elston-Stewart algorithm [5]. Computation time is highly dependent on the pedigree structure and the number of untyped founders.

A reduction in the number of likelihood function evaluations would allow more samples, larger and more complex pedigrees, or a greater density of markers to be analyzed in a reasonable amount of time. Although the GPS method [2] was more robust than the older line search method implemented in all previous versions of ILINK, we decided that the number of likelihood evaluations might be reduced by using instead a newer algorithm known to outperform GPS in some other optimization problems.

Mesh Adaptive Direct Search (MADS) [6] is a framework for a class of derivative-free algorithms designed to supersede the GPS method. MADS is conceptually similar to GPS, but uses a richer set of search directions, resulting in better theoretical convergence properties. The NOMAD software package [7] is a high-quality, C++ open-source implementation of MADS algorithms in use in universities and companies around the world. [8–11] NOMAD is robust [12] and has a wide range of functionality, including handling

of general nonlinear constraints, biobjective optimization, parallelism, and the restriction of variables to integer or boolean values. [13]

We describe PSEUDOMARKER 2.0, which uses a customized version of ILINK that uses NOMAD to maximize likelihoods. We show that NOMAD is more effective at finding optima than GPS, while requiring fewer evaluations of the likelihood function.

Implementation
PSEUDOMARKER

PSEUDOMARKER uses parametric inheritance models and exact likelihood computations to evaluate the evidence for linkage and/or LD between a putative trait locus and a set of genotyped markers. When applying extreme parametric models, it yields statistics that are stochastically equivalent to several popular model-free methods if applied to simple family structures [14], for instance mother-father-child triads, case-control samples, or affected sib-pairs. PSEUDOMARKER, however, has substantial advantages over the simpler nonparametric methods when analyzing more complex family structures [4].

PSEUDOMARKER uses likelihood ratio tests to compare four models describing all possible combinations of having or not having linkage and having or not having LD. Marker allele frequencies are parameters of all four likelihood functions. For likelihoods allowing for LD, the marker allele frequencies are allowed to vary conditional on which trait-locus allele is on the same haplotype. For likelihoods allowing for linkage, the probability with which recombination occurs between trait and marker loci (the recombination fraction) is a parameter. For each likelihood function, all parameters are estimated jointly.

Estimating the parameters is a nonlinear constrained optimization problem. ILINK uses the pedigree structure, genomic data and the inheritance model to compute each likelihood function exactly as a nonlinear function of its free parameters. Marker allele frequencies and conditional allele frequencies are probabilities, and as such are constrained to lie between 0 and 1. Each set of frequencies must also sum to 1. The recombination fraction, if a parameter, is constrained to lie between the 0 and 0.5, because larger values of the recombination fraction are not biologically meaningful; a recombination fraction of 0.5 between two loci indicates that the loci segregate independently.

NOMAD

NOMAD [7] implements several variants of the MADS framework for constrained derivative-free optimization. In its usual mode, it searches for an optimum by generating trial points along orthogonal directions starting from the incumbent best solution [15]. The set of directions used in this step is far richer than the set of directions searched by GPS; formally, the set of normalized directions is dense in the unit sphere. The use of such a rich set of search directions ensures stronger theoretical convergence properties, and leads to a more efficient algorithm in practice [6]. The MADS framework is flexible enough to allow the use of heuristics that investigate additional trial points to improve practical convergence. Heuristics available in NOMAD include Variable Neighborhood Search (VNS) metaheuristic [16] and the construction and exploration of quadratic models of the objective function and of the constraints. [12]. The VNS metaheuristic was not used in our tests, but quadratic models are used by NOMAD in its default mode, and were used in our tests.

To optimize likelihoods, NOMAD proposes to ILINK values for its free parameters, trial points in the MADS framework. ILINK attempts to evaluate the likelihood function at these trial points. NOMAD explicitly handles bound constaints, and so will not, for instance, suggest a negative probability. The constraints that allele frequencies sum to 1 was handled by another of NOMAD's features, the extreme barrier approach. For any set of marker allele frequencies, one frequency may represented implicitly, its value obtained by subtracting the sum of the other frequencies from 1. NOMAD is not aware of the implicit frequencies. For a trial point suggested by NOMAD, it is possible for an implicit frequency to have an infeasible value: a negative value or a value greater than one. In such a circumstance, the extreme barrier takes effect. ILINK informs NOMAD that the trial point is infeasible, and NOMAD ignores the point, effectively treating it as if it had an infinitly bad objective value.

ILINK was modified substantially to use NOMAD instead of GPS. The interface between PSEUDOMARKER and ILINK was modified to enable better performance, but these changes do not affect PSEUDOMARKER usage. NOMAD was run in a mode that uses $2n$ orthogonal search directions, where n represents the number of optimization variables. NOMAD was stopped when the minimum poll size, a NOMAD parameter, was less than 1e-4, indicating that for the next set of trial points, the largest change to any parameter to the likelihood functions would be at most 1e-4.

4

Computational experiments

Table 1 gives a brief summary of the 12 datasets that we analyzed in this project. The datasets contained both biallelic markers and multiallelic microsatellites. Table 2 shows pedigree statistics of the data sets; more detailed statistics are shown in Additional file 1: Tables S1–S3. Pedigree, phenotype, and marker statistics were computed using PedStats [17]. Simulated genotype data were generated using a modified version of SLINK [18, 19]; parameters used for the simulations are shown in Additional file 2: Tables S4 and S5.

Test problems were selected to include difficult cases, including such factors as real life pedigree structures, realistic amounts of missing data, and large multi-generational families. The real datasets were from Finnish gene mapping studies on which TH and JDT were collaborators [20, 21], while the simulated data sets were generated as part of the Ph.D. dissertation of TH, some of which have been analyzed in prior publications [1, 22].

Some data sets were observed to present difficult maximization problems for the GPS while the previous version of the PSEUDOMARKER package was being developed. The x.linked test set [23] was particularly interesting because it was x-linked, had multiple alleles, and most of the data were triads, and still maximization was quite time-consuming.

All 12 sets were analyzed under assumptions of both the dominant and recessive extreme inheritance models described in [14] and all four likelihood functions used by PSEUDOMARKER, testing for linkage and/or LD. Six were also analyzed under more biologically plausible inheritance models. We optimized likelihoods using either GPS as previously described [1] or NOMAD [7].

Results and Discussion

The numbers of likelihood function evaluations for each test set, summed over all markers, all models, and all maximized likelihood functions, are shown in Table 3. NOMAD is superior in terms of function evaluations to GPS on all test sets. As we discuss below, NOMAD is invoked somewhat differently from GPS on the same optimization problems, which contributes to the improvement.

In preliminary tests, we observed NOMAD was more robust than GPS in finding an optimum (data not shown). Because of this observation, we experimented with invoking NOMAD less often. For GPS, it was

often helpful to retry a given optimization problem, using the solution previously returned from GPS as the new starting point because that would sometimes lead to the identification of a better likelihood value. The purpose of these restarts is to encourage convergence to a global optimum, and to reduce the probability that GPS would stall at a non-optimal point. The restarts were unnecessary with NOMAD. Nor was it helpful to start NOMAD at several different initial estimates, as was done with GPS. The counts in Table 3 are counts for invoking NOMAD once to solve each optimization problem, whereas GPS was invoked as described in [1].

Despite the fewer calls to the optimization algorithm, the optimum returned by NOMAD was usually better than the one from GPS. Of the 288 optimization problems we tried based on the 12 test sets, NOMAD found an assignment to the variables that yielded a log likelihood that was at least 0.005 worse than the value reported by GPS only seven times (see Table 4 and Additional file 3: Table S6). In contrast, NOMAD reported 68 objective values better by at least 0.005 than the values reported by GPS. We considered differences less than 0.005 in the log likelihood to be insubstantial, as such differences would change log of the likelihood ratio by at most 0.01. NOMAD returned answers with objective value more than 0.5 better than GPS 21 times, with the largest difference being 28, a shockingly large value. In contrast, the most GPS improved the objective value over NOMAD was 0.1.

Conclusions

The new PSEUDOMARKER 2.0 has been released (see Availability and Requirements) and it uses NOMAD [7] to maximize likelihoods. The new version usually provides better or comparable answers, while using far fewer evaluations of the likelihood functions. Several of the most prominent developers of pedigree analysis methods recognized decades ago that the optimization problems that arise in genetic analysis of pedigrees can be difficult to solve and can benefit from new methods [24–26]. We have shown in this study that MADS methods are more effective than previous methods on the optimization problems that arise in usage of PSEUDOMARKER [1]. We suggest therefore, that NOMAD may be useful in other mathematical optimization problems arising in genetics.

Availability and Requirements

Project name: PSEUDOMARKER 2.0

Project home page: http://www.helsinki.fi/~tsjuntun/pseudomarker/2.0/

Operating system(s): GNU/Linux Intel 64-bit archetecture

Programming language: C and C++

Other requirements: none

License: PSEUDOMARKER is a binary distribution with registration required. (Referees may obtain a copy of PSEUDOMARKER from this site without registration.) NOMAD is distributed with PSEUDOMARKER under terms of the LGPL 3.0.

Any restrictions to use by non-academics: no

Acknowledgements

List of abbreviations

LD: linkage disequilibrium; GPS: generalized pattern search; MADS: mesh adaptive direct search

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

EMG modified FASTLINK to use NOMAD, performed the experiments and wrote the first draft of the manuscript. EMG and TH updated PSEUDOMARKER to use NOMAD. TH collected test sets, wrote test scripts, and edited the manuscript. CA and SLD helped integrate NOMAD with FASTLINK, suggested tests, and edited the manuscript. JDT suggested tests and edited the manuscript. AAS concieved the project, helped integrate NOMAD with FASTLINK, suggested tests, and edited the manuscript. All authors read and approved the final manuscript.

Authors' information

References

1. Hiekkalinna T, Schäffer AA, Lambert B, Norrgrann P, Göring HHH, Terwilliger JD: **PSEUDOMARKER: a powerful program for joint linkage and/or linkage disequilibrium analysis on mixtures of singletons and related individuals**. *Hum. Hered.* 2011, **71**(4):256–266.

2. Dennis JE Jr, Torczon V: **Direct search methods on parallel machines**. *SIAM J. Optim.* 1991, **1**(4):448–474.

3. Cottingham RW, Idury RM, Schäffer AA: **Faster sequential genetic linkage computations**. *Am. J. Hum. Genet.* 1993, **53**:252–263.

4. Hiekkalinna T, Göring HHH, Lambert B, Weiss KM, Norrgrann P, Schäffer AA, Terwilliger JD: **On the statistical properties of family-based association tests in datasets containing both pedigrees and unrelated case-control samples**. *Eur. J. Hum. Genet.* 2012, **20**(2):217–223.

5. Elston RC, Stewart J: **A general model for the genetic analysis of pedigree data**. *Hum. Hered.* 1971, **21**(6):523–542.

6. Audet C, Dennis JE Jr: **Mesh adaptive direct search algorithms for constrained optimization**. *SIAM J. Optim.* 2006, **17**:188–217.

7. Le Digabel S: **Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm**. *ACM Trans. Math. Softw.* 2011, **37**(4):1–15.

8. Stracquadanio G, Romano V, Nicosia G: **Semiconductor device design using the BiMADS algorithm**. *J. Comput. Phys.* 2013, **242**:304–320.

9. Torres R, Bès C, Chaptal J, Hiriart-Urruty JB: **Optimal, environmentally-friendly departure procedures for civil aircraft**. *J. Aircraft* 2011, **48**:11–22.

10. Aasi J, Abadie J, Abbott BP, Abbott R, Abbott TD, Abernathy M, Accadia T, Acernese F, Adams C, Adams T, Addesso P, Adhikari R, Affeldt C, Agathos M, Agatsuma K, Ajith P, Allen B, Allocca A, Amador Ceron E, Amariutei D, Anderson SB, Anderson WG, Arai K, Araya MC, Ast S, Aston SM, Astone P, Atkinson D, Aufmuth P, et al.: **Einstein@Home all-sky search for periodic gravitational waves in LIGO S5 data**. *Phys. Rev. D* 2013, **87**:042001.

11. Alarie S, Audet C, Garnier V, Le Digabel S, Leclaire LA: **Snow water equivalent estimation using blackbox optimization**. *Pac. J. Optim.* 2013, **9**:1–21.

12. Conn AR, Le Digabel S: **Use of quadratic models with mesh adaptive direct search for constrained black box optimization**. *Optim. Methods Softw.* 2013, **28**:139–158.

13. Audet C, Le Digabel S, Tribes C: **NOMAD user guide**. Tech. Rep. G-2009-37, Les cahiers du GERAD 2009, [http://www.gerad.ca/NOMAD/Downloads/user_guide.pdf].

14. Göring HHH, Terwilliger JD: **Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified**. *Am. J. Hum. Genet.* 2000, **66**(4):1310–1327.

15. Abramson M, Audet C, Dennis JE Jr, Le Digabel S: **OrthoMADS: A deterministic MADS instance with orthogonal directions**. *SIAM J. Optim.* 2009, **20**(2):948–966.

16. Audet C, Béchard V, Le Digabel S: **Nonsmooth optimization through mesh adaptive direct search and variable neighborhood search**. *J. Global Optim.* 2008, **41**(2):299–318.

17. Wigginton JE, Abecasis GR: **PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data**. *Bioinformatics* 2005, **21**(16):3445–3447.

18. Weeks DE, Ott J, Lathrop GM: **SLINK: a general simulation program for linkage analysis**. *Am. J. Hum. Genet.* 1990, **47**(Suppl):A204.

19. Schäffer AA, Lemire M, Ott J, Lathrop GM, Weeks DE: **Coordinated conditional simulation with SLINK and SUP of many markers linked or associated to a trait in large pedigrees**. *Hum. Hered.* 2011, **71**(2):126–134.

20. Wessman M, Kallela M, Kaunisto MA, Marttila P, Sobel E, Hartiala J, Oswell G, Leal SM, Papp JC, Hämäläinen E, Broas P, Joslyn G, Hovatta I, Hiekkalinna T, Kaprio J, Ott J, Cantor RM, Zwart JA, Ilmavirta M, Havanka H, Färkkilä M, Peltonen L, Palotie A: **A susceptibility locus for migraine with aura, on chromosome 4q24**. *Am. J. Hum. Genet.* 2002, **70**(3):652–662.

21. Ekelund J, Hovatta I, Parker A, Paunio T, Varilo T, Martin R, Suhonen J, Ellonen P, Chan G, Sinsheimer JS, Sobel E, Juvonen H, Arajärvi R, Partonen T, Suvisaari J, Lönnqvist J, Meyer J, Peltonen L: **Chromosome 1 loci in Finnish schizophrenia families**. *Hum. Mol. Genet.* 2001, **10**(15):1611–1617.

22. Hiekkalinna T: **On the superior power of likelihood-based linkage disequilibrium mapping in large multiplex families compared to population based case-control designs**. *PhD thesis*, University of Helsinki, Helsinki, Finland 2012.

23. Karjalainen MK, Huusko JM, Ulvila J, Sotkasiira J, Luukkonen A, Teramo K, Plunkett J, Anttila V, Palotie A, Haataja R, Muglia LJ, Hallman M: **A potential novel spontaneous preterm birth gene, AR, identified by linkage and association analysis of X chromosomal markers**. *PLoS ONE* 2012, **7**(12):e51378.

24. Elston RC, Lange E, Namboodiri KK: **Age trends in human chiasma frequencies and recombination fractions. II. Method for analyzing recombination fractions and applications to the ABO:nail-patella linkage**. *Am. J. Hum. Genet.* 1976, **28**:69–76.

25. Ott J: **Counting methods (EM algorithm) in human pedigree analysis: Linkage and segregation analysis**. *Ann. Hum. Genet.* 1977, **40**(4):443–454.

26. Weeks DE, Lange K: **Trials, tribulations, and triumphs of the EM algorithm in pedigree analysis**. *IMA J. Math. Appl. Med. Biol.* 1989, **6**(4):209–232.

27. Pajukanta P, Terwilliger JD, Perola M, Hiekkalinna T, Nuotio I, Ellonen P, Parkkonen M, Hartiala J, Ylitalo K, Pihlajamäki J, Porkka K, Laakso M, Viikari J, Ehnholm C, Taskinen MR, Peltonen L: **Genomewide scan for familial combined hyperlipidemia genes in finnish families, suggesting multiple susceptibility loci influencing triglyceride, cholesterol, and apolipoprotein B levels**. *Am. J. Hum. Genet.* 1999, **64**(5):1453–1463.

28. Kaunisto MA, Tikka PJ, Kallela M, Leal SM, Papp JC, Korhonen A, Hämäläinen E, Harno H, Havanka H, Nissilä M, Säkö E, Ilmavirta M, Kaprio J, Färkkilä M, Ophoff RA, Palotie A, Wessman M: **Chromosome 19p13 loci in Finnish migraine with aura families**. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 2005, **132B**:85–89.

29. Tikka-Kleemola P, Artto V, Vepsäläinen S, Sobel EM, Räty S, Kaunisto MA, Anttila V, Hämäläinen E, Sumelahti ML, Ilmavirta M, Färkkilä M, Kallela M, Palotie A, Wessman M: **A visual migraine aura locus maps to 9q21-q22**. *Neurology* 2010, **74**(15):1171–1177.

Tables and captions

Table 1 - Summary of all data sets

Test set	Description	Reference
fin1	Familial combined hyperlipidemia pedigrees from Finland	Pajukanta *et al.* [27]
fin2	Migraine pedigrees from Finland	Wessman *et al.* [20], Kaunisto *et al.* [28], Hiekkalinna *et al.* [4]
fin3	A sub set of the Migraine families (different phenotype and genotyped individuals than on data set fin2)	Tikka-Kleemola *et al.* [29]
fin4	Schizophrenia families from Finland	Ekelund *et al.* [21], Hiekkalinna *et al.* [4]
fin5	Same as fin1, but with multiallelic markers	
fin6	Same as fin1, but with highly polymorphic marker	
x.linked	Extended pedigrees and triads from northern Finland with real X-chromosomal marker data	Karjalainen *et al.* [23]
100sibs	Artificial sib-pair pedigrees	Hiekkalinna [22]
100sibs.c	Artificial sib-pair pedigrees with additional cases	Hiekkalinna [22]
100sibs.cc	Artificial sib-pair pedigrees with additional cases and controls	Hiekkalinna [22]
mixed	Various size artificial pedigrees (triads, sib-pairs, and extended pedigrees)	Hiekkalinna [22]
noparents	Artificial affected sib-pairs with no parental genotypes	Hiekkalinna [22]

Table 2 - Data set properties

Data set	Pedigrees	Average pedigree size	Singleton Cases	Singleton Controls	Number of Markers	Maxium Alleles/Marker
fin1	61	15.33	200	200	3	2
fin2	84	13.08	200	200	3	2
fin3	37	13.24	100	100	4	4
fin4	438	5.79	0	199	3	2
fin5	61	15.33	200	200	4	8
fin6	61	15.33	200	200	1	18
x.linked	482	3.17	112	203	1	20
100sibs	100	4.00	0	0	1	3
100sibs.c	100	4.00	200	0	1	3
100sibs.cc	100	4.00	200	200	1	3
mixed	180	5.22	0	50	6	3
noparents	200	4.50	100	100	2	4

Table 3 - Number of function evaluations used by GPS and NOMAD

Test Set	GPS	NOMAD	Test Set	GPS	NOMAD
fin1	7,650	3,342	100sibs	10,003	3,933
fin2	7,430	3,341	100sibs.c	10,891	3,240
fin3	81,887	10,765	100sibs.cc	7,137	2,811
fin4	8,460	3,250	mixed	39,522	12,278
fin5	83,272	32,662	noparents	34,590	9,143
fin6	284,069	96,626	x.linked	470,517	140,986

Table 4 - Changes in objective function

Count of changes in the objective function more extreme than the indicated number. Positive changes indicate that NOMAD found the better objective value.

Data set	≤ -0.5	≤ -0.05	≤ -0.005	≥ 0.005	≥ 0.05	≥ 0.5
fin3	0	0	0	23	15	8
fin4	0	1	1	0	0	0
fin5	0	0	0	20	13	7
fin6	0	0	0	8	4	2
x.linked	0	2	6	4	4	1
mixed	0	0	0	2	2	2
noparents	0	0	0	11	7	1

Additional Files
Additional file 1 — test_set_statistics.pdf

Contains Tables S1–S3, showing statistical information about the test sets.

Additional file 2 — simulation_parameters.pdf

Contains Tables S4–S5, showing parameters used to generate the simulated genotypes in the test sets.

Additional file 3 — objective_values.pdf

Contains Table S6, showing differences in the objective value computed by GPS and by NOMAD.

Additional files provided with this submission:

Additional file 1: test_set_statistics.pdf, 113K
http://www.biomedcentral.com/imedia/4056793701060486/supp1.pdf
Additional file 2: simulation_parameters.pdf, 213K
http://www.biomedcentral.com/imedia/1957551141060486/supp2.pdf
Additional file 3: objective_values.pdf, 131K
http://www.biomedcentral.com/imedia/1052376538106048/supp3.pdf
Additional file 4: Reference PDF.pdf, 144K
http://www.biomedcentral.com/imedia/1108730868106052/supp4.pdf